Statistical modeling for prediction of diabetes in Malaysians

Amatul Zehra [1], Tuty Asmawaty Abdul Kadir[1], M.A.M. Aznan[2], Riaz-ul Haq [1], Gran Badshah [1]

[1.] Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang, 26300, Kuantan, Malaysia
[2.] Kulliyah of Medicine, International Islamic University Malaysia, P.O Box 141, Kuantan, Pahang 25710, Malaysia
amatulzehrak@gmail.com

Abstract: Type II Diabetes Mellitus is one of the silent killer diseases worldwide. According to the World Health Organization, 347 million people are suffering from diabetes throughout the world. To overcome the sharp rise in the disease, various diagnostic or prediction models were developed through various techniques such as artificial intelligence, classification and clustering, pattern recognition and statistical methods. The study led to the related open issues of identifying the need of a relation between the major factors that lead to the development of diabetes. This is possible by investigating the links found between the independent and dependant variables in the dataset. This paper investigates the effect of binary logistic regression applied on a dataset. The results show that the most effective method was the enter method which gave a prediction accuracy of almost 93%.
[Zehra A, Kadir TAA, Aznan MAM, Haq R, Badshah G. **Statistical modeling for prediction of diabetes in Malaysians.** *Life Sci J* 2018;15(6):76-80]. ISSN: 1097-8135 (Print) / ISSN: 2372-613X (Online). http://www.lifesciencesite.com. 9. doi:10.7537/marslsj150618.09.

Keywords: Type II Diabetes Mellitus; diabetes prediction, binary logistic regression

1. Introduction

Diabetes Mellitus has become a common health problem nowadays, which would affect people and lead to various disablements like cardio vascular diseases, visual impairments, leg amputations and renal failure if diagnosis is not done in the right time (World Health Organization, 2013). Diabetes can affect people due to the lack of insulin in the blood. Insulin is a natural hormone secreted by the pancreas, which acts as a key to unlock the body cells so that sugar, starch and food molecules can be absorbed and hence be utilized by the cells to generate energy required for daily life. Insulin deficiency is due to either of the two conditions. First is when the pancreas does not produce insulin at all. This leads to type I diabetes mellitus (T1DM) which is usually found by birth. Second state is when the body does not respond correctly to the insulin produced by the pancreas and hence the glucose that is consumed by the person is locked inside the blood instead of entering into the cells of the body. This ineffective insulin leads to type II diabetes mellitus (T2DM). Among these, type I diabetes is usually diagnosed in children and type II is the most common form which affects adults and is preventable (Pobi, 2006).

2. Diabetes-a global threat

The International Diabetes Federation has estimated an alarming rise in the number of diabetics by the year 2030 (International Diabetes Federation, 2012). A sharp rise in diabetics has been observed in Asian region with 138 million Asians including 14.9% Malaysians (The Malay Mail, 2010). From 1996 to 2006, the number of diabetics in Malaysia had increased by almost 80% and reached to 1.4 million adults above the age of 30. Among those, almost 36% were undiagnosed; resulting in complications that required more intensive medical care, putting great strain on the existing overstretched health services (The Star, 2010).

The need for avoidance and better management of type II diabetes has been an important issue since ages. Medical practitioners and researchers have investigated and continue to find solutions to overcome this disease. Various researches and studies are done on predicting the blood glucose levels for type II diabetes patients for a short term. Most of the predictions helped to decide the diet control and physical activities in order to maintain a healthy life (Lauritzen, 2011). A comparison study in which three algorithms of neural network are compared shows that the Bayesian Regulation gives the highest prediction accuracy of 88.8% (Sapon, 2011). A study was conducted to develop and validate a predictive equation to check for undiagnosed diabetes using American subjects. The predictive equation was validated through Egyptian subjects using multiple logistic regression method. The resultant equation gave a positive predicted value of 63% (Tabaei, 2002). In a different study, 25,639 patients were studied for a period of 5 years to assess whether a risk score comprising only physical parameters was effective to identify the risk of diabetes (Rahman, 2008). The results showed that those in the top

quintile of risk were 22 times more likely to develop diabetes than those in the bottom quintile (odds ratio 22.3; 95% CI: 11.0–45.4). In all, 54% of all clinically incident cases occurred in individuals in the top quintile of risk (risk score > 0.37). The area under the ROC was 74.5%. In another study, a population-based risk prediction tool named '(Diabetes Population Risk Tool (DPoRT)' was developed using national survey data to predict 9-year risk for diabetes (Rosella, 2010). In another research, the ability of various metabolic syndrome criteria in the prediction of diabetes was compared with the ability to determine whether various proposed modifications to the National Cholesterol Education program (NCEP) metabolic syndrome definition improved predictive capacity (Hanley, 2005). The results showed that modifications or additions to the NCEP metabolic syndrome definition had limited impact on the prediction of diabetes. In a similar research, the predictive value of different parameters responsible for the incidence of type 2 diabetes mellitus (T2DM) in subjects with metabolic syndrome was investigated (Ozery-Flato, 2013). The results showed that fasting plasma glucose (FPG), body mass index (BMI), and glycosylated hemoglobin can be used to predict diabetes onset with a high level of accuracy and each was shown to have a cumulative predictive value.

This paper focuses to investigate the major factors that affect the onset of diabetes in Malaysian population. Based on those factors, we developed the empirical equation that can be used to predict diabetes on a different sample. As compared to the previous work, this paper discusses the room for improvement in the classification equation for predicting diabetes. We aimed to study type II diabetes because this type can be prevented by adopting proactive measures. This paper focuses to apply the binary logistic regression using its various methods on a dataset provided by the Ministry of Health, Malaysia. Eventually, the model would be able to answer the need for significant and urgent requirement to: (i) stop sharp rise in diabetes, (ii) grow public health awareness, and (iii) prevent the onset of this disease.

3. Material and Methods

To develop a predictive equation for diabetes, a dataset of 28,498 cases was used which was provided by the Ministry of Health, Malaysia. This dataset is part of the outcome of The National Health and Morbidity Survey 2011 (NHMS 2011) (Fadhli, 2013). The parameters which were provided in the dataset were individual code, enumeration block, post stratification weight, strata, state locality, state, age, gender, race group, physical activity level, blood glucose level, fasting status, diabetes status, cholesterol level, average systolic blood pressure, average diastolic blood pressure, waist and body mass index (BMI) status. For the purpose of analysis, we found out that the parameters titled individual code, enumeration block, post stratification weight, strata, state locality and fasting status did not have a direct impact on diabetes. After removing these variables, we were left with a total of eleven variables that were important for the study. Out of these eleven variables, 5 were categorical and 6 were continuous variables. Also, there were surplus cases in the dataset such as individuals less than 18 years of age. After deleting those cases and the missing values, we were left with 14,767 cases. A number of 1240 cases were already known diabetic and the rest 13,527 cases were non-diabetic to their knowledge. The number of males in the dataset was 7011 and 7756 were females. The states of Malaysia that were under study were Johor, Kedah, Kelantan, Melaka, Negeri Sembilan, Pahang, Penang, Perak, Perlis, Selangor, Terengganu, Sabah/Labuan, Sarawak, Wilayah Persekutuan Kuala Lumpur, Wilayah Persekutuan Labuan or Wilayah Persekutuan Putrajaya. The ranges of ages of the cases were between 18 years to 107 years. The subjects belonged to the races of Malays, Chinese, Indians, other Bumiputeras or others. There were 9822 cases who claimed that they were active and the rest 4945 were inactive. The blood glucose levels ranged between 2.00 mmol/L and 32.20 mmol/L. The ranges of cholesterol levels ranged between 2.59 mmol/L and 10.35 mmol/L. The average systolic and diastolic blood pressure ranges were between 69 to 244 mmHg and 35 to 170 mmHg respectively. The waist circumference ranged between 41 to 166 cm. The body mass index (BMI) variable had four categories namely underweight, normal, overweight or obese.

The dependent variable in this dataset is the 'diabetes status' variable which states whether the patient is diabetic or not. The outcome of this variable is Yes or No, where Yes=1 and No=0. All the other variables are considered as independent variables or covariates which have an impact on the dependent variable. The demographic characteristics of the study sample are shown in Table 1.

Table 1: Demographic characteristics of the study sample

Variable	Characteristics
N	14,767
State	1298 from Johor, 926 from Kedah, 1009 from Kelantan, 972 from Melaka, 848 from Negeri Sembilan, 546 from Pahang, 1009 from Penang, 908 from Perak, 852 from Perlis, 2038 from Selangor, 800 from Terengganu, 1678 from Sabah/Labuan, 938 from Sarawak, 475 from WP Kuala Lumpur, 470 from WP Putrajaya
Age	18 ~ 107 years
Gender (male/female)	7011 males / 7756 females
Race	8318 Malays, 2834 Chinese, 1205 Indians, 1459 Other Bumiputera or 951 Others
Physical Activity Level (active/inactive)	9822 active / 4945 inactive
Blood glucose level (mmol/L)	2.00 ~ 32.20 mmol/L
Known diabetes mellitus (yes/no)	1240 diabetics / 13527 non-diabetics
Cholesterol level (mmol/L)	2.59 ~ 10.35 mmol/L
Average systolic blood pressure (mmHg)	69 ~ 244 mmHg
Average diastolic blood pressure (mmHg)	35 ~ 170 mmHg
Waist circumference (cm)	41 ~ 166 cm
BMI Status	1118 Underweight, 6693 Normal, 4518 Overweight or 2438 Obese

4. Binary logistic regression

Binary logistic regression, often referred to simply as logistic regression, is used to predict a categorical variable from a set of predictor variables (Osborne, 2007). Logistic regression is often chosen if the predictor variables are a mix of continuous and categorical variables and/or if they are not nicely distributed. It makes no assumptions about the distributions of the predictor variables. Logistic regression has been especially popular with medical research in which the dependent variable is whether or not a patient has a disease.

For a logistic regression, the predicted dependent variable is a function of the probability that a particular subject will be in one of the categories. This can be written as:

$$P - \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

where P is the probability of a 1, e is the base of the natural logarithm (about 2.718) and α and β are the parameters of the model. The value of α yields P when x is zero, and β indicates how the probability of a 1 changes when x changes by a single unit. Because the relation between x and P is nonlinear, β does not have as straightforward an interpretation in this model as it does in ordinary linear regression (Tabachnick, 2007).

The logistic regression method was applied using SPSS software. SPSS is predictive analytics software that can predict with confidence what will happen next so that smarter decisions can be made, problems be solved and outcomes be improved. We chose to use SPSS because the aim of this study was to develop an empirical equation using existing data and based on that equation we can predict the future patients who are at risk.

5. Results

A direct logistic regression was performed on diabetes status as outcome and eleven predictors: state, age, gender, race group, physical activity level, blood glucose level, diabetes status, cholesterol level, average systolic blood pressure, average diastolic blood pressure, waist and body mass index (BMI) status. Analysis was performed using SPSS. Using the logistic regression analysis on the existing data, the outcome of the analysis is shown in Table 2. We have used the 'enter' method which forces all of the independent variables into the model. The enter method is the most common and recommended method in statistics.

Table 2: Regression coefficients, Wald statistics, odd ratios and confidence intervals.

Variables in the Equation					
	B	**S.E.**	**Wald**	**Sig.**	**Exp (B)**
State	-.035	.009	14.670	.000	.966
Age	.057	.003	383.276	.000	1.059
Gender	.246	.077	10.091	.001	1.279
Race	-.006	.035	.026	.873	.994
Physical Activity	.171	.075	5.212	.022	1.186
Blood Glucose level	.343	.011	1021.803	.000	1.409
Cholesterol	-.292	.032	80.810	.000	.747
Systolic blood pressure	.004	.002	2.809	.094	1.004
Diastolic blood pressure	-.008	.004	4.218	.040	.992
Waist circumference	.024	.004	34.236	.000	1.024
BMI	.248	.064	15.135	.000	1.281
Constant	-9.18	.414	491.400	.000	.000

6. Discussions

A test of the full model with all eleven predictors against a constant-only model was statistically significant, χ^2 (7, N=14,737) = 23.19, <.001, indicating that the predictors, as a set, reliably distinguished between diabetics and non-diabetics. Classification was impressive, with 98% of the non-diabetic and 30.3% of the diabetics correctly predicted, for an overall success rate of 92.9%.

Table 2 shows regression coefficients, Wald statistics, odds ratios, and 95% confidence intervals for odds ratios for each of the eleven predictors. Direct logistic regression was performed to assess the impact of a number of factors on the likelihood of getting diabetes. The full model containing all predictors was statistically significant, $\chi 2$ (11, N = 14,767) = 1021, p <.001, indicating that the model was able to distinguish between diabetics and non-diabetics. The model as a whole explained between 17.8% (Cox and Snell R square) and 40.6% (Nagelkerke R squared) of the variance in diabetics, and correctly classified 92.9% of cases. As shown in Table 2, seven of the independent variables made a unique statistically significant contribution to the model (state, age, gender, blood glucose level, cholesterol, waist circumference and BMI). The strongest predictor of diabetes was blood glucose level, recording an odds ratio of 1.409. In addition to the blood glucose level, all the odds ratios that were above 1 were age, gender, systolic blood pressure, waist circumference and BMI. This implies that one unit increase in these variables will lead to the increase in the number of diabetics multiplied by the respective odds. The odds ratios of less than 1 are the variables of state, race, cholesterol and diastolic blood pressure. This indicates that a unit increase in these variables will lead to a decrease in the number of diabetics by the respective odds. The odds ratio for gender indicates that when holding all other variables constant, a man is 1.27 times more likely to get diabetes than a woman. Inverting the odds ratio for waist circumference reveals that for each one point increase there is a 1.024 times increase of the odds that the participant will get diabetes.

7. Conclusion

Diabetes is one of the leading causes of deaths worldwide. In order to control the disease, there is a need to predict the onset of diabetes through models and applications. In this paper, we conducted a binary logistic regression analysis on the data of National Health and Morbidity Survey 2011, Malaysia. We found out that there are a few parameters that are significant in prediction of diabetes with a success rate of 92.9 %. In future, the same data can be used to predict other diseases that have are directly related to diabetes and its complications.

Acknowledgements

The authors would like to thank the Director General of Health, Malaysia for his permission to use the data from the National Health and Morbidity Survey 2011 and to publish this paper.

Corresponding Author:
Amatul Zehra
Faculty of Computer Systems & Software Engineering
Universiti Malaysia Pahang, 26300,
Kuantan, Malaysia
E-mail: amatulzehrak@gmail.com

References

1. World Health Organization, http://www.who.int/topics/diabetes_mellitus/en/, (Last access date: 30th September 2012).

2. Pobi S., Hall LO., Predicting juvenile diabetes from clinical test results. International Joint Conference on Neural Networks (IJCNN). 2006; 2159 – 65.
3. International Diabetes Federation, http://www.idf.org/diabetesatlas/5e/regional-overviews, (Last access date: 30th September 2012).
4. 'Sharp rise of diabetics in Asia', The Malay Mail, 3rd December 2010, (Last access date: 13th October 2011).
5. 'Alarming rise in number of diabetics in Malaysia', The Star, 11th January 2010, (Last access date: 12th October 2011).
6. Lauritzen J.N., Arsand E., Vuurden K.V., Bellika J.G., Hejlesen O.K. and Hartvigsen G., Towards a mobile solution for predicting illness in type 1 diabetes mellitus: Development of a prediction model for detecting risk of illness in type 1 diabetes prior to symptom onset. 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronics Systems Technology (Wireless VITAE), 2011; 1-5.
7. Sapon M.A, Ismail K. and Zainudin S., Prediction of Diabetes by using Artificial Neural Network, International Conference on Circuits, System and Simulation, 2011; 299-303.
8. Tabaei B.P and Herman W.H., A Multivariate Logistic Regression Equation to Screen for Diabetes, *Diabetes Care,* 2002; (25): 1999-2003 .
9. Rahman M., Simmons R.K, Harding A.H, Wareham N.J and Griffin S.J, A simple risk score identifies individuals at high risk of developing Type 2 diabetes: a prospective cohort study, *Family Practice Advance Access*, 2008;191-196 .
10. Rosella L.C, Manuel D.G, Burchill C. and Stukel T.A, A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT), *J Epidemiol Community Health*, 2011; (65): 613-620 .
11. Hanley A.J.G, Karter A.J, Williams K., Festa A., D'Agostino R.B., Wagenknecht L.E and Haffner S.M., Prediction of Type 2 Diabetes Mellitus With Alternative Definitions of the Metabolic Syndrome: The Insulin Resistance Atherosclerosis Study, *Circulation*, 2005; (112): 3713-3721.
12. Ozery-Flato M., Parush N., El-Hay T., Visockienė Z., Ryliškytė L., Badarienė J., Solovjova S., Kovaitė M., Navickas R. and Laucevičius A., Predictive models for type 2 diabetes onset in middle-aged subjects with the metabolic syndrome, *Diabetology & Metabolic Syndrome*, 2013; (36): 1-9.
13. Fadhli Y., Azahadi O., Noor Ani A., Balkish M.N, Ahmad Jessree K. and Tahir A., Approaches in Methodology of a Population-Based Study in Malaysia: The National Health and Morbidity Survey 2011 (NHMS 2011), *Malaysian Journal of Medicine and Health Sciences,*2013; (9): 25-33.
14. Osborne J.W., Best Practices in Quantitative Methods. 2007 SAGE Publications, Inc.
15. Tabachnick B.G. and Fidell L.S., Using Mulivariate Statistics, 2007 5th Ed. Pearson Education Inc.

6/22/2018